Analysis between sequential feature engineering with statistical features for malware behaviour discovery

TU Delft

1 Background

Malware Packet-sequence Clustering and Analysis (MalPaCA)[2]

- Sequential clustering of malware behaviour
- Uses 4 features: Packet size, Delay, Source Port & Destination Port
- Non-intrusive features, respects user privacy
- Hierarchical Density-Based Spatial Clustering (HDBScan) as clustering algorithm

IoT-23[1] **Data set**

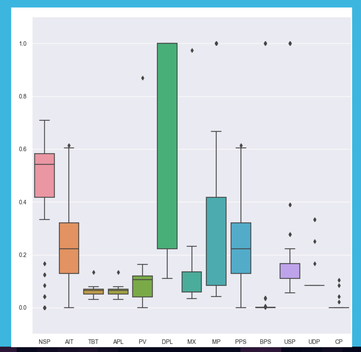
The dataset consists of 20 infected IoT devices and 3 benign cosisting of 100GB of network traffic PCAP files .
A Selection of data set for experiments was made to ensure a variance of different malicious labels

Goal

- Define statistical features set for uni-directonal flows
- Compare Sequential features to Statistical features
- Recommend improvements for MalPaCA

2 Statistical Features

Feature	Description
NSP	Number of small packet
AIT	Average arrival time of packets
TBT	Total number of transmitted bytes
APL	Average payload packet length
PV	Standard deviation of payload packet length
DPL	The total of number of different packet sizes
MX	Size of largest packet
MP	The number of maximum packets
PPS	Number of packets per second
BPS	Average bits-per-second
USP	Total number of unique Source ports
UDP	Total number of unique Destination ports
CP	Common ports in Source & Destination ports



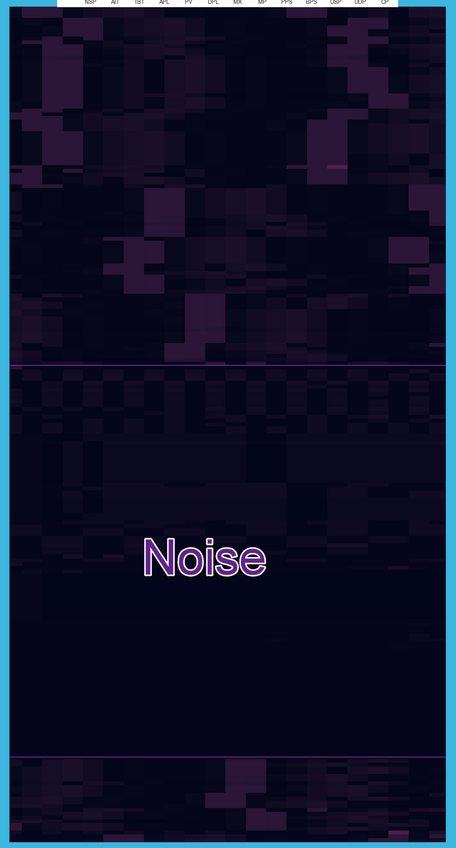
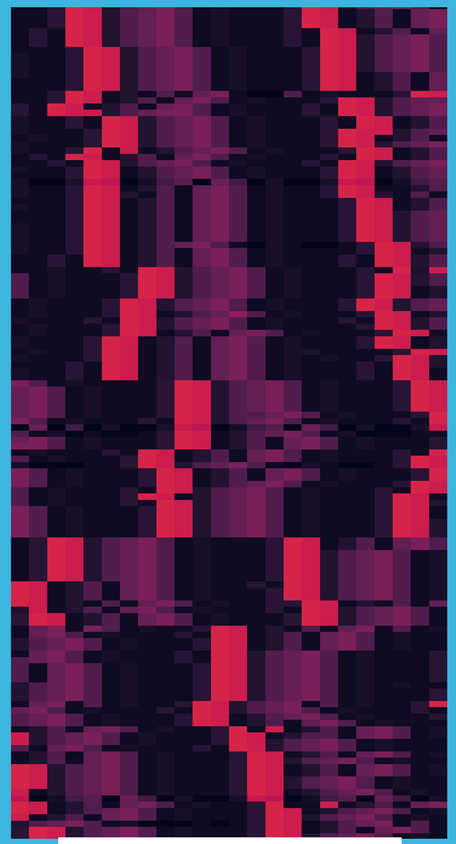


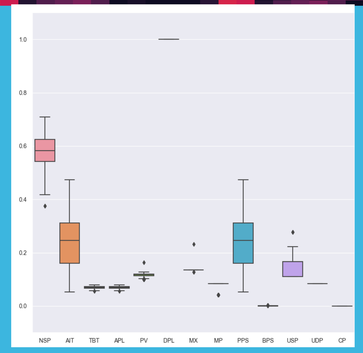
Figure 1: Heatmap of byte size 'Attack' cluster generated by Sequential features with statistical feature distribution. (Y= Connections, X = Packet)

Sequential

Statistical

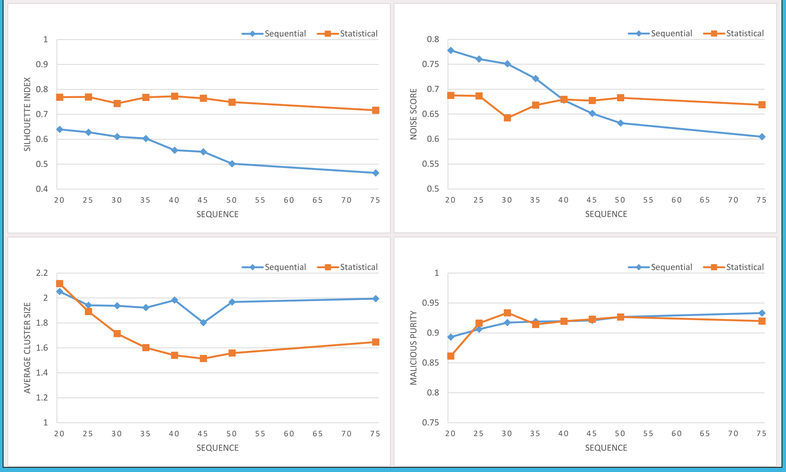
Figure 2: Heatmap of byte size 'Attack' cluster generated by Statistical features with statistical feature distribution. (Y= Connections, X = Packet)





3 Result

Average over 100 runs with variance in input data



4 Conclusion

Sequential and statistical features create distinguishable clusters containing a distinct network behaviour with very similar purity & malicious purity.

Sequential features excel low sequence length with an overal lower noise score between lenghts 20-40.

Sequential features are better at generalizing clusters with the same clustering parameters (Higher average cluster size)

Clusters generated with statistical features have higher cohesion and are further seperated from each other over all sequence lengths.

MalPaCA could fully use all statistical features for longer sequence lengths while using sequential for short sequences.

The feature sets are not exclusive and could be used togheter.

References

[1] Sebastian Garcia, Agustin Parmisano, and Maria Jose Erquiaga. "IoT-23: A Labeled Dataset with Malicious and Benign IoT Network Traffic." Zenodo, January 20, 2020. https://doi.org/10.5281/zenodo.4743746.

[2] Nadeem, Azqa, Christian Hammerschmidt, Carlos H. Ganan, and Sicco Verwer. "Beyond Labeling: Using Clustering to Build Network Behavioral Profiles of Malware Families." In Malware Analysis Using Artificial Intelligence and Deep Learning, edited by Mark Stamp, Mamoun Alazab, and Andrii Shalaginov, 381–409. Cham: Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-62582-5_15.

Contact

Researcher: Mikhail Epifanov
Supervisors: Azqa Nadeem
Sicco Verwer

1 July 2021